VALUE IN HEALTH 15 (2012) 504–513

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Preference-Based Assessments

Condition-Specific Preference-Based Measures: Benefit or Burden?

Matthijs M. Versteegh, MA, BSc, Annemieke Leunis, MSc, Carin A. Uyl-de Groot, PhD, Elly A. Stolk, PhD*

iMTA/iBMG, Institute of Health Policy and Management/Institute for Medical Technology Assessment, Erasmus University of Rotterdam, Rotterdam, The Netherlands

ABSTRACT

Objectives: Some argue that generic preference-based measures (PBMs) are not sensitive to certain disease-specific improvements. To overcome this problem, new condition-specific PBMs (CS-PBMs) are being developed, but it is not yet clear how such measures compare with existing generic PBMs. **Method:** We generated CS-PBMs from three condition-specific questionnaires (Health Assessment Questionnaire for arthritis, Quality of Life Questionnaire for Cancer 30 for cancer, and Multiple Sclerosis Impact Scale 29 for multiple sclerosis). First, the questionnaires were reduced in content, and then, a time trade-off study was conducted in the general public (N = 402) to obtain weights associated with the dimensions and levels of the new questionnaire. Finally, we compared utilities obtained by using the CS-PBMs with utilities obtained by using the EuroQol five-dimensional (EQ-5D) questionnaire in four data sets. **Results:** Utility values generated by the CS-PBMs were higher than those of the EQ-5D questionnaire. The Health Assessment Questionnaire–based measure for arthritis proved to be insensitive to comorbidities. Measures based on the Multiple Sclerosis Impact Scale 29 and the Quality of Life Questionnaire for Cancer 30 discriminated comorbidities and side effect equally well as the EQ-5D questionnaire and were more sensitive than the EQ-5D questionnaire for mild impairments. **Conclusions:** The introduction of PBMs that are specific to a certain disease may have the merit of sensitivity to disease-specific effects of interventions. That gain, however, is traded off to the loss of comparability of utility values and, in some cases, insensitivity to side effects and comorbidity. The use of a CS-PBM for cost-utility analysis is warranted only under strict conditions.

Keywords: patient-reported outcome measures, preference-based measures, time trade-off, utility.

Introduction

A preferred method for generating the quality adjustment required for the computation of quality-adjusted life-years is through generic preference-based measures (PBMs) such as the EuroQol five-dimensional (EQ-5D) questionnaire [1] or the health utilities index [2]. Some argue that such generic PBMs are not sensitive to certain disease-specific improvements. Consequently, the existing PBMs may not always be the best tool to assess the effect of an intervention. To overcome this problem, new condition-specific PBMs (CS-PBMs) have been developed, for example, for asthma [3] and urinary incontinence [4]. Not much is known, however, about how these new instruments compare with generic instruments such as the EQ-5D questionnaire. It is feared that using CS-PBMs may lead to the exaggeration of health problems due to a focusing effect, render comparison of utility values impossible, because utilities are derived from different PBMs, and may be insensitive to comorbidities [5,6]. Evidence, however, is scarce. In this study, three CS-PBMs were developed for the purpose of exploring these and other issues, one for arthritis (based on the Health Assessment Questionnaire [HAQ]), one for multiple sclerosis (MS) (based on the Multiple Sclerosis Impact Scale 29 [MSIS-29]), and one for cancer (based on the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire for Cancer 30 [QOL-C30]).

A PBM is a questionnaire with a scoring function to weight the responses according to preferences for certain health conditions over others. These preference weights are elicited in studies where respondents are asked to express their preference for a health state, for instance, using time trade-off (TTO) or standard gamble. Existing generic PBMs such as the EQ-5D questionnaire and the health utilities index were developed to have a standardized tool to measure the health-related quality of life for the quality adjustment part of the quality-adjusted life-year. These generic preference-based instruments aim to measure the quality of life on a sufficient degree of generality to allow comparisons across conditions. For these instruments, the key trade-off is between generality of the included health dimensions to allow cross-disease comparisons and sensitivity to pick up (relevant) treatment effects [5]. The EQ-5D questionnaire, for example, consists of five items with three levels measuring mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. The choice to include only these basic dimensions of health ensures the level of generality that is required for comparison across diseases at the potential cost of losing sensitivity for disease-specific complaints. For

* *Address correspondence to:* Matthijs M. Versteegh, iMTA/iBMG, Institute of Health Policy and Management/Institute for Medical Technology Assessment, Erasmus University of Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands.

E-mail: versteegh@bmg.eur.nl.

1098-3015/$36.00 – see front matter

doi:10.1016/j.jval.2011.12.003

example, the view is widely held that the EQ-5D questionnaire is not an appropriate measure to assess the quality of life of patients with sensory problems (bad eyesight or hearing problems), because sensory problems are beyond the scope of health defined by dimensions of the EQ-5D questionnaire [7]. Another perceived problem of the EQ-5D questionnaire is that very mild conditions cannot be adequately assessed by using only three levels of impairment because of low ceiling sensitivity [8,9].

The increased use of economic evaluations by health authorities seems to have created a sense of urgency within the health assessment community to deal with the shortcomings of generic PBMs. In recent years, new CS-PBMs have emerged for which the development was motivated by either the absence of generic PBMs in a specific context or the judgment that generic PBMs would not be appropriate for a condition. Contrary to generic instruments, a CS-PBM contains dimensions specifically targeted at the affected population. In terms of the trade-off mentioned above, these instruments are expected to demonstrate superior sensitivity to specific diseases, although this may come at the cost of comparability of utility values across conditions. Because of the difference in the scope of different instruments, utility values derived from a CS-PBM may not be comparable with those derived from a generic instrument, even though they seem to lie on the same 0-to-1 scale. Although the development of CS-PBMs is valuable for research purpose, for example, to better investigate the shortcomings of generic PBMs, there is concern about the application of CS-PBMs in economic evaluations. Unfortunately, empirically founded guidance on how and when to apply CS-PBMs is absent.

There has been little reflection so far on the comparability of the obtained quality-of-life weights to those obtained from generic PBMs. Specific issues in comparability are described in a recent expert editorial [5]. First, CS-PBMs may cause an exaggeration of health problems (reflected by low utility values) due to focusing effects. When the health states in a preference elicitation study consist of a set of disease-related items, rather than general items of health-related quality of life, the context of the valuation is narrower, possibly leading to lower utility values. The logic behind this hypothesis is that narrow-focused items may seem less important when presented in a wider context of general health (e.g., having a cold may seem less severe when presented alongside problems with mobility), but may seem quite problematic when presented separately. This may result in a downward bias on preference values when compared with generic PBMs. Second, a CS-PBM might have difficulty capturing comorbidities as the focus is on disease-related items. This may result in an upward bias on utility values. Furthermore, developing a CS-PBM is not a clear-cut exercise. Researchers face many decisions, such as the reduction of items in a questionnaire, the selection of health states (how many and which?) that have to be valued to develop a PBM [7], on the valuation method (e.g., TTO or standard gamble?), and on the modeling approach. How these decisions are dealt with may differ per study, which decreases comparability.

The primary aim of this article was to provide empirical evidence about the comparability of CS-PBMs and generic PBMs. To do so, three CS-PBMs were developed from existing questionnaires. The values generated by these CS-PBMs were then compared with EQ-5D questionnaire values for the same patient samples. By providing empirical evidence we hope to provide a better understanding of the effects of using CS-PBMs and contribute to the development of guidance for their use. This is important, because it can be expected that in the nearby future more cost-utility analyses will contain utilities based on condition-specific measures.

Methods

Questionnaires for CS-PBM development

The CS-PBMs were developed from the HAQ [10], the MSIS-29 [11], and the QLQ-C30 [12]. These instruments were selected on the basis of expert advice and commonality of use within clinical settings. The HAQ is a widely used questionnaire in rheumatology to measure functional abilities by using 20 items with four levels spread across eight domains (dressing, rising, eating, walking, hygiene, reach, grip, and usual activities). The scale has been shown to be unidimensional [13]. The MSIS-29 measures the impact of MS on a physical and psychological dimension. Dimensionality of the subscales has been confirmed by using Rasch analysis [14]. The QLQ-C30 (version 2) is a cancer-specific questionnaire consisting of 30 items. These items cover five functional scales, nine symptom scales, and a global health status scale. These questionnaires were chosen because they differ in scope and because EQ-5D questionnaire data were available for the purpose of comparing results. For MSIS-29, it has been shown that the physical scale is better capable of discriminating among subcategories of the clinically assessed Expanded Disability Status Scale (EDSS) than is the EQ-5D questionnaire [15]. There was no evidence known to us on a lack of responsiveness of the EQ-5D questionnaire or the superiority of the condition-specific measures HAQ and QLQ-C30 in arthritis and cancer, respectively.

Reducing the content of the questionnaires

Developing a PBM from an existing questionnaire does not lead to an entirely new instrument but attaches weights to some of the items of the existing questionnaire. Such an approach generally requires a method to reduce the questionnaire content because only a limited number of items can be valued in a preference elicitation study [7]. Typically, only a fraction of the total amount of all theoretically possible health states is valued. The values for the remainder of the health states are estimated through modeling techniques.

The optimal number of items in a health state was considered to be in the order of five to nine items, because more items may cause difficulties for respondents in the valuation study [7]. The HAQ, MSIS-29, and QLQ-C30, respectively, contain 20, 29, and 30 items, and so reduction of content was required. Relevant and well-functioning items from the questionnaires were selected by using the following criteria proposed by others [16]: 1) the item had to fit the Rasch model, 2) the item had to meet basic psychometric criteria, and 3) the selected items had to be approved by a clinical expert. Four data sets were available for these analyses: the Rotterdam Early Arthritis CoHort for the HAQ (N = 738), the Multiple Sclerosis Risk Sharing Scheme Monitoring Study (N = 1295) for the MSIS-29, and the Haemato Oncology Foundation for Adults in the Netherlands 24 (pooled N = 716) and Haemato Oncology Foundation for Adults in the Netherlands 25 (pooled N = 789) trials for the QLQ-C30. The data set characteristics are described in detail in Versteegh et al. [15]. A set of a priori criteria was used to determine which items were suitable for the health state description [16,17]. Because it was expected that neither of these criteria could be sufficient on its own, the three criteria were employed "side by side" (i.e., no hierarchical order).

Criterion 1: Fit to the Rasch model

Rasch analysis was used to test the psychometric validity of a scale and to identify well-functioning items. The Rasch model assumes that the probability of scoring level λ on item i is a logistic function of the relative distance between the item location (how much disability it represents) and the respondent location (how disabled the patient is) [18].

The main performance criteria within the Rasch model were whether the item 1) has ordered thresholds (having more of the latent trait θ results in endorsing a higher level answer category [19]), 2) fits the Rasch model (fit residual <2.5 and nonsignificant bonferroni-adjusted probability), 3) combined scale fits the Rasch model (described by a nonsignificant item-trait interaction chi-square probability [19]), and 4) shows no differential item functioning. After each single scale amendment the analysis was rerun for the remaining items. Rasch analysis was performed on the dimensional structure originally suggested by the questionnaires.

Criterion 2: Psychometric properties

Psychometric criteria were laid alongside the Rasch results to come to a final selection of items amenable for valuation. The functioning of the items was tested by investigating the loading of items on factors identified by factor analysis; missing data; internal consistency of items with its scale score; distribution of the responses on an item including floor and ceiling effects; and regression coefficients between a general health indicator and an item. Psychometric analyses were applied to the full data sets.

Criterion 3: Expert opinion

The selected items from the questionnaires were presented to experts in the respective fields. Experts from the Erasmus Medical Centre and the VU Amsterdam Medical Centre were consulted to gain insight into important aspects of the disease and to evaluate the result of the previous selection process.

Health state selection

Even after data reduction the selected set can still generate an enormous amount of possible health states; therefore, a fractional factorial design was favored over a full factorial design. The QLQ design was a level-balanced design, meaning that all levels of each item occurred with the same frequency. Within the balanced design, health states covered the entire spectrum of severity, measured by averaging the item levels of a health state. For the MSIS-29 and the HAQ, items and levels were selected with an orthogonal main effects plan (OMEP) as is applied in other studies [20,21] to ensure zero statistical correlation between the attributes. The set was complemented with a selection of the most observed health states (four or more observations) over the severity range of the questionnaire. TTO values estimated with additive main-effect models (one based on the OMEP states and one based on the OMEP and the most observed states) were compared with the observed TTO values of the most occurring states by using standard predictive performance measures such as mean absolute error (MAE) to see whether the addition of these states led to improved prediction of the most frequently occurring states.

The final design was blocked. In such a design respondents value a number of health states that belong to the same "block." The mean severity of the combination of items in a block was similar and measured through summing the level scores of the items in a block.

Health state valuation with time trade-off method

The preferences of a sample of the general public were elicited through a TTO exercise for each of the selected health states of the questionnaires. To optimize comparability to generic PBMs, the CS-PBM health states were valued with the same TTO protocol, the same computer-assisted personal interview tool, the same procedure to measure states "worse than dead," and the same rescore procedure of negative values (negative TTO scores were rescaled to have a range between -1 and 0 with ($-$t/$-$x$-$t) as was adopted in the Dutch EQ-5D questionnaire valuation study) [22]. Unlike the Dutch EQ-5D questionnaire valuation study, this study was performed in group sessions, which has previously been shown to produce comparable TTO results [23].

The TTO exercise was self-administered through a digital tool for TTO elicitation (computer-assisted personal interviews) in groups with about 12 to 25 respondents per session. Each session was supervised by three to four researchers to offer assistance if needed. Prior to the task, respondents received 30 minutes of instructions by researchers M.V. or A.L. including examples of the TTO computer program projected on a large screen. The task was piloted by M.V. and A.L. in a sample of 18 respondents to ensure the introduction, the computer program, and the organization of the task were feasible.

The three questionnaires were presented separately in the TTO exercise and in all possible orders (e.g., first HAQ, then MSIS, then QLQ). Within the TTO exercises, health states were presented random to individuals.

Respondents

Respondents were selected by a marketing agency that requited a sample resembling the Dutch general population in age, gender, and education. Respondents were approached by phone and asked whether they were interested in contributing to a task to value descriptions of health states. Respondents received a financial reward of €35 upon completion of the three TTO exercises. Respondents were removed from the analyses when the results indicated they valued the majority of logically worse states higher than logically better states in a set (i.e., HAQ state 11112 is logically better than HAQ state 14444).

Modeling of the TTO values

Once the TTO study had been performed, the preference values observed for the selected health states were used to estimate

Table 1 – Items selected for the TTO valuation exercise.

HAQ	MSIS-29	QLQ-C30
• HAQ1 Stand up from a straight chair • HAQ2 Walk outdoors on flat ground • HAQ3 Get on / off toilet • HAQ4 Reach and get down a 5-pound object (such as a bag of sugar) from just above your head • HAQ5 Open car doors	• MSIS1 Problems with your balance • MSIS2 Being clumsy • MSIS3 Limitations in your social and leisure activities at home • MSIS4 Difficulties using your hands in everyday tasks • MSIS5 Having to cut down the amount of time you spent on work or other daily activities • MSIS6 Feeling mentally fatigued • MSIS7 Feeling irritable, impatient or short tempered • MSIS8 Problems concentrating	• QLQ1 Trouble taking a long walk • QLQ2 Limited in doing either your work or other daily activities • QLQ3 Have you had pain • QLQ4 Have you felt nauseated • QLQ5 Were you tired • QLQ6 Difficulty in concentrating on things • QLQ7 Did you worry • QLQ8 Has your physical condition or >medical treatment interfered with your social activities

Table 2 – TTO study design following item selection.

	HAQ	MSIS	QLQ
Number of items	5	8	8
Total number of health states to be valued	56	100	105
States identified by OMEP (used in study after fold-over)	15 (30)	32 (64)	N/A
Number of most occurring states included	26	36	N/A
Number of states valued by one individual (total = 33)	8	10	15
Number of blocks*	7	10	7
Expected number of observations per health state (N = 400/number of blocks)	57	40	57

HAQ, Health Assessment Questionnaire; MSIS, Multiple Sclerosis Impact Scale; N/A, not applicable/available; OMEP, orthogonal main effects plan; QLQ, Quality of Life Questionnaire; TTO; time trade-off.
* One block consist of a number of states, and all the states in one block are valued by one individual.

values for all potential health states through statistical modeling. Because individuals value more than one health state there are multiple observations for each individual. Random effects models were estimated to assess how the predictors (the items and their levels) influence the dependent variable (the mean observed TTO value). In these random effects models, the item levels were treated as dummy variables with dummy coding. The constant term was treated as an additional decrement for having any item level other than the base case ("no problems"), which is similar to the EQ-5D questionnaire model. The values predicted by this random effects model will be referred to as the PBM results (e.g., HAQ-PBM). Models were required to have significant predictors and worse scores on the levels ought to be represented by larger utility decrements. Model performance was assessed by comparing the MAE of observed and predicted values. Models were estimated until meeting those criteria. Only the most parsimonious models are presented. To keep optimal comparability between the developed CS-PBMs, models were estimated from the items only, without interaction effects or a "worst-value" dummy variable, which is 1 for every item on the lowest level. Interaction effects were not estimated because the study design was a main effects design.

Hypotheses and analyses

To investigate the properties of PBMs developed from existing questionnaires, several hypotheses were tested. First, it was tested whether the TTO values could be successfully modeled. For HAQ and MSIS, the TTO random effects models were fitted on both the full data set (including "most observed" health states) and on the subset consisting of health states originating from the OMEP. This was done to test whether an OMEP alone is sufficient to estimate the utility values of the most occurring health states. Second, it was investigated whether CS-PBMs yielded lower mean utility values than did a generic measure, which was hypothesized to reflect that a downward bias on utility values resulting from a focusing effect might outweigh the upward bias on utility values resulting from a narrower scope of the CS-PBM. Third, it was tested with Wilcoxon rank-sum tests, to account for the non-normal distribution of utility values, whether the developed CS-PBMs had a more narrow focus and were therefore less sensitive to comorbidities (in arthritis and MS data sets) or side effects (non-Hodgkin's lymphoma data set) than the EQ-5D questionnaire. Side effects had World Health Organization performance status 2 or higher, representing the inability to carry out work activities due to the condition. Fourth, we assessed discriminative properties of the new measures by using clinical indicators. For arthritis, the Disease Activity Score-28 was used, which is based on a count of tender joints and the erythrocyte sedimentation rate. It distinguished between high, moderate, and low disease activity and remission. For MS, the EDSS was used, which, when rounded to integers, distinguishes 11 categories of increasing disability. For cancer, we used the World Health Organization performance status score (or Eastern Cooperative Oncology Group), which distinguishes six categories, from 0 to 6 (death). Lastly, responsiveness was measured in the cancer population by using effect size (Cohen's *d*) and mean change in the cancer population, for which follow-up measurements were available in the data set.

All results were compared with utilities of the Dutch EQ-5D questionnaire tariff [22].

Software

For Rasch analysis, the RUMM 2020 software (Rumm Laboratory Pty Ltd., Perth, Western Australia) was used. Psychometric analysis was performed in SPSS 17.0 (SPSS Inc., Chicago, IL, USA), and all hypothesis testing and modeling efforts were performed in STATA 11.0 (StataCorp. 2009, College Station, TX, USA).

Table 3 – Respondent characteristics.

	TTO study sample	Dutch population norms*
N	402	–
Gender, M/F (%)	46/54	49.5/50.5
Age (y)		
Mean (SD)	45 (15.5)	40.1
Min–max	15–76	–
Age group (y) (%)		
<20	4.8	23.7
20–40	37.6	25.3
40–65	46.9	35.7
65–80	10.4	11.4
>80	0.3	3.9
Education (%)		
High	34	27
Medium	35	31
Low	25	33
Missing/else	6	9
Mean (SD) time to complete TTO (min)		
HAQ 8 states	8 (4.6)	–
MSIS 10 states	10 (5.8)	–
QLQ 15 states	12.7 (5.8)	–

HAQ, Health Assessment Questionnaire; MSIS, Multiple Sclerosis Impact Scale; N/A, not applicable/available; OMEP, orthogonal main effects plan; QLQ, Quality of Life Questionnaire; TTO; time trade-off.
* Statistics Netherlands, 2009 figures.

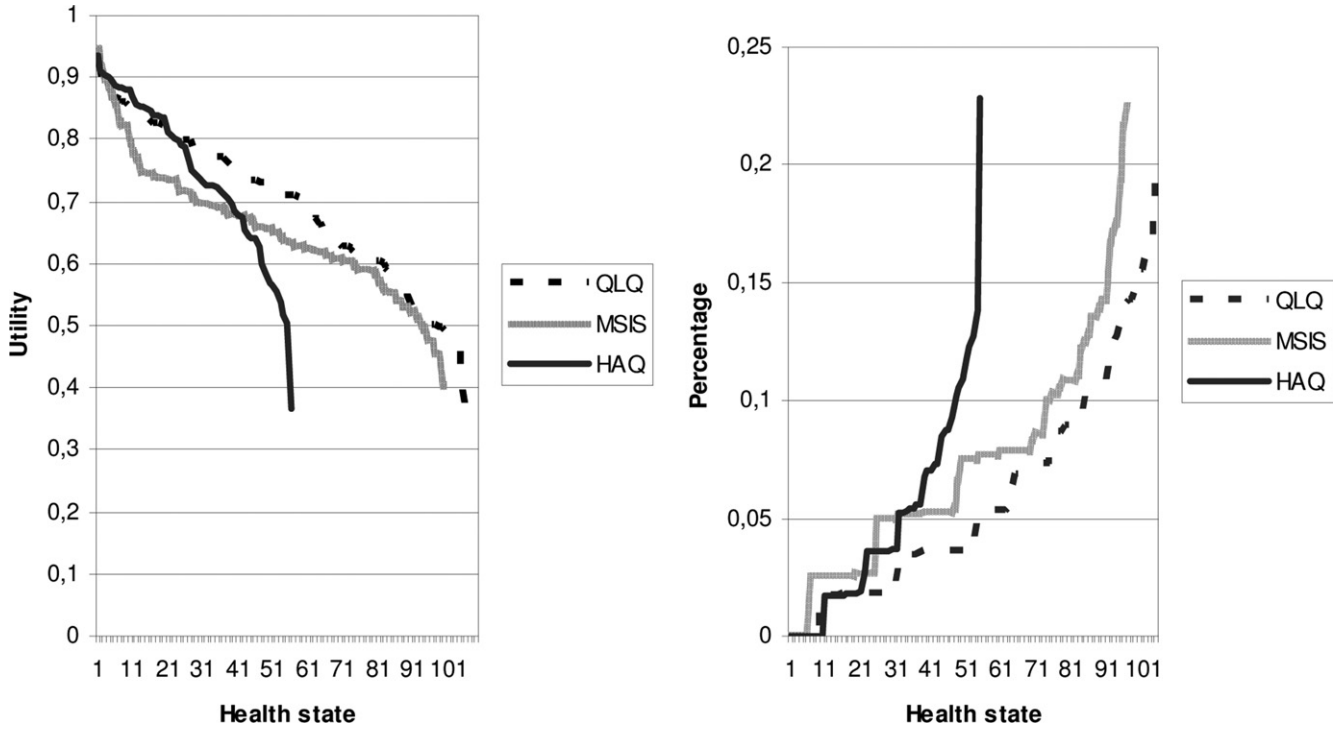

Fig. 1 – (A) Mean utility values of health states. (B) Percentage of respondents who classified a state as worse than dead. HAQ, Health Assessment Questionnaire; MSIS, Multiple Sclerosis Impact Scale; QLQ, Quality of Life Questionnaire.

Results

Item and level selection

The selected items per questionnaire are presented in Table 1, and all met the criteria of the Rasch analysis, psychometric analysis, and expert opinion. The full results of the selection of items and levels are described in Appendix A. A table of the results of the Rasch analysis is presented in Appendix B (see Appendices in online Supplemental Materials found at doi: 10.1016/j.jval.2011.12.003).

Resulting study design

Given the many items and levels in the study, we chose a fractional factorial blocked design. Health states were presented in blocks, so that one individual values one block containing several health states. The design is listed in Table 2.

Data quality

Four hundred two respondents participated in the computer-assisted TTO study and resembled the Dutch population (Table 3). Respondents were excluded from the analyses if they had valued the majority of logically better states lower than logically worse states in one block (8 exclusions for HAQ, 17 for MSIS, and 7 for QLQ). Average time to value one health state in the TTO exercise was about 1 minute. Total time per block was highest for QLQ (15 health states, about 12 minutes), followed by MSIS (10 health states, 10 minutes) and HAQ (8 health states, 8 minutes). Although two separate researchers took turns in holding the introductory talks, this did not bias the TTO responses (Wilcoxon rank-sum test: $P > 0.05$). On average, women had higher utility values (t–test: $P < 0.00$) for all three questionnaires.

The mean utility of the health states and the percentage of responses indicating a state to be worse than dead are presented in Figure 1A, B.

Table 4 – Final random effects model characteristics.

	HAQ-PBM*	MSIS-PBM†	MSIS-PBM*	QLQ-PBM
Random effects mean models				
*R-square**	0.94	0.68	0.78	0.88
MAE	0.028	0.034	0.04	0.033
MAE most observed states	0.022	0.057	0.043	–
Illogical sign or order of variables	0	0	0	0
Insignificant predictors	0	0	0	0
Possible range	0.32–1	0.40–1	0.42–1	0.34–1

HAQ-PBM, Health Assessment Questionnaire-preference-based measure; MAE, mean absolute error; MSIS-PBM, Multiple Sclerosis Impact Scale-preference-based measure; QLQ-PBM, Quality of Life Questionnaire-preference-based measure.

* Model based on states from the orthogonal design and the most observed states.

† Model based on states from the orthogonal design.

Table 5 – Coefficients of random effects models with TTO value as dependent variable.

HAQ-PBM			MSIS-PMB*			QLQ-PMB		
	Coefficient	SE		Coefficient	SE		Coefficient	SE
haq1_2	**−0.005**	0.001	ms1_2	**−0.016**	0.003	qlq1_2	**−0.027**	0.001
haq1_3	**−0.031**	0.002	ms1_3	**−0.043**	0.003	qlq2_2	**−0.020**	0.002
haq1_4	**−0.121**	0.002	ms1_4	**−0.089**	0.003	qlq2_3	**−0.047**	0.002
haq2_2	**−0.029**	0.001	ms2_2	**−0.018**	0.003	qlq2_4	**−0.068**	0.002
haq2_3	**−0.091**	0.002	ms2_3	**−0.047**	0.003	qlq3_3	**−0.079**	0.002
haq2_4	**−0.144**	0.002	ms2_4	**−0.047**	0.003	qlq3_4	**−0.213**	0.001
haq3_2	**−0.042**	0.001	ms3_3	**−0.055**	0.002	qlq4_2	**−0.018**	0.002
haq3_3	**−0.055**	0.002	ms3_4	**−0.071**	0.002	qlq4_3	**−0.055**	0.002
haq3_4	**−0.213**	0.002	ms4_2	**−0.061**	0.002	qlq4_4	**−0.089**	0.002
haq4_2	**−0.022**	0.001	ms4_3	**−0.101**	0.003	qlq5_2	**−0.021**	0.002
haq4_3	**−0.041**	0.002	ms4_4	**−0.108**	0.003	qlq5_3	**−0.031**	0.002
haq4_4	**−0.074**	0.002	ms5_2	**−0.032**	0.003	qlq5_4	**−0.037**	0.002
haq5_2	**−0.016**	0.001	ms5_3_4†	**−0.057**	0.002	qlq6_2	**−0.004**	0.002
haq5_3	**−0.038**	0.002	ms6_2	**−0.020**	0.003	qlq6_3	**−0.039**	0.002
haq5_4	**−0.044**	0.002	ms6_3	**−0.035**	0.003	qlq6_4	**−0.052**	0.002
Constant	**0.918**	0.002	ms6_4	**−0.059**	0.003	qlq7_3	**−0.009**	0.002
			ms7_3	**−0.024**	0.002	qlq7_4	**−0.047**	0.002
			ms7_4	**−0.038**	0.002	qlq8_2	**−0.008**	0.002
			ms8_2	**−0.037**	0.003	qlq8_3	**−0.041**	0.002
			ms8_3	**−0.059**	0.003	qlq8_4	**−0.060**	0.002
			ms8_4	**−0.073**	0.003	Constant	**0.944**	0.002
			Constant	**0.959**	0.005			

HAQ-PBM, Health Assessment Questionnaire-preference-based measure; MSIS-PBM, Multiple Sclerosis Impact Scale-preference-based measure; QLQ-PMB, Quality of Life Questionnaire-preference-based measure; TTO, time trade-off.

* MSIS model with most observed health states included.

† Both ms5_3 and ms5_4 have the same decrement.

Bold indicates all coefficients were significant at $P<0.05$.

Modeling

TTO values were modeled for each of the three questionnaires with random effects mean prediction models. For the HAQ, using only the OMEP-based health states had too much variation in TTO scores between respondents to estimate a model with significant predictors and logical negative signs for each of the dummy variables (increasing negative decrements per item level of severity). Estimating the model on all the available data (thus including the "most observed" states) yielded a well-functioning mean prediction model. The prefinal MSIS-29 model had insignificant predictors for three variables MSIS3; the prefinal QLQ-C30 model had insignificant predictors for two variables. In all instances, merging the levels with the adjacent categories resolved the problem. Model characteristics are summarized in Table 4, and full models are presented in Table 5.

When the MSIS-29 prediction model was based on all the states (thus including the most observed states), the prediction error for the most observed states was reduced (MAE = 0.043 compared with MAE = 0.057). When the MSIS-29 TTO values were modeled without the "most observed" states, the utility values were generally higher, which caused the utility values of some of the "most observed" states to be overestimated (Fig. 2).

Comparability of mean utility values

In the four data sets, the developed CS-PBMs based on the models presented in Table 5 produced a higher mean utility score for patients than did the EQ-5D questionnaire (Table 6). Especially, the HAQ-PBM (mean = 0.91) had a much higher mean utility value than did the EQ-5D questionnaire (mean = 0.68). Furthermore, the difference between the mean EQ-5D questionnaire score in arthritis and the mean EQ-5D questionnaire score in MS was 0.06 while the difference between the mean HAQ-PBM and the MSIS-PBM score was 0.24. The QLQ-C30-PBM–based utility values had the highest correlation with EQ-5D questionnaire utility values. Both the MSIS-PBM and the QLQ-PBM, however, have increased sensitivity compared with the EQ-5D questionnaire. Where the EQ-5D questionnaire scores full health (a utility value of 1), the MSIS-PBM and the QLQ-PBM report decrements in utility for 99 and 185 patients, respectively (Table 7).

Comorbidities and side effects

The HAQ-PBM could not discriminate between patients with and without comorbidity (other vascular disorders and psychiatric disorders) when the EQ-5D questionnaire could (Table 8). For arthritis pa-

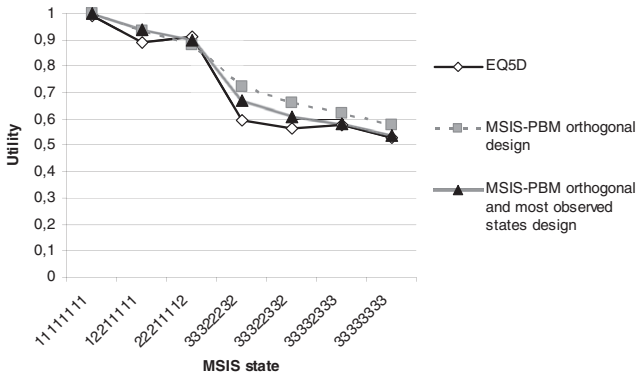

Fig. 2 – Utility values for most observed MSIS states in patient data set. EQ-5D, EuroQol five-dimensional; MSIS, Multiple Sclerosis Impact Scale; MSIS-PBM, Multiple Sclerosis Impact Scale-preference-based measure.

Table 6 – Comparison of utility values derived from the new PBMs with the EQ-5D questionnaire and the SF-6D.

	HAQ	MSIS	QLQ_MM	QLQ_NH
N	738	1295	716	789
Mean utility (SD) [range]				
EQ-5D questionnaire	0.68 (0.23) [−0.134 to 1]	0.62 (0.26) [−0.22 to 1]	0.74 (0.21) [−0.058 to 1]	0.73 (0.26) [−0.33 to 1]
SF-6D	0.66 (0.10) [0.37–1]	–	–	–
PBM*	–	0.69 (0.13) [0.40–1]	0.84 (0.09) [0.44–1]	0.82 (0.11) [0.34–1]
PBM†	0.91 (0.09) [0.57–1]	0.67 (0.14) [0.42–1]	–	
Intraclass correlations (ICC)				
EQ-5D questionnaire-PBM*†	0.45	0.62	0.64	0.67

EQ-5D, EuroQol five-dimensional; HAQ, Health Assessment Questionnaire; MM, multiple myeloma; MSIS, Multiple Sclerosis Impact Scale; NH, non-Hodgkin; PBM, preference-based measure; QLQ, Quality of Life Questionnaire; SF-6D, six-dimensional health state short form (derived from short form 36 health survey).

* Model based on states from the orthogonal or balanced design.

† Model based on states from the orthogonal design and the most observed states.

tients with diabetes, hypercholesterolemia, or thyroid disease, the HAQ-PBM showed higher utility values for individuals with the disorder while the EQ-5D questionnaire signaled the expected direction of differences. The MSIS-PBM also showed higher utilities for patients with asthma and high blood pressure (rather than without), but this was concordant with the differences indicated by the EQ-5D questionnaire. Both the MSIS-PBM and the EQ-5D questionnaire picked up significant differences between MS patients with and without depression ($P < 0.05$). In the non-Hodgkin's lymphoma data set, patients with side effects and infections as a result of treatment had lower ($P < 0.05$) utility values in both the EQ-5D questionnaire and the QLQ-C30 than did patient without side effects and infections, except for hair loss. All significant differences were at least half a SD except for comorbidity "depression" in the MS data set and "other side effects" in the non-Hodgkin's lymphoma data set.

Discriminative ability and responsiveness

Utilities of all instruments decreased with an increase in severity as assessed by the clinical indicator (Table 9). The utilities of the HAQ-PBM, however, failed to distinguish between low and moderate disease activities. The EQ-5D questionnaire did so accurately. As has previously been shown, the EQ-5D questionnaire was unable to distinguish between categories 3, 4, and 5 on the EDSS [15]. This signifies the inability of the EQ-5D questionnaire to distinguish between fully ambulatory patients with MS (EDSS 3) and patients whose disability is severe enough to impair full daily working activities (EDSS 5). The MSIS-PBM, of which the physical scale was known to be sensitive to changes between level 3, 4, and 5, did pick up the deterioration in health. Neither the QLQ-PBM nor the EQ-5D questionnaire adequately reflected the deterioration between level 0 and level 1 of the World Health Organization performance status.

The QLQ-C30 was, in terms of effect size measured with Cohen's *d*, at times more and at times less sensitive to changes over time (Table 10). However, the absolute differences indicated that even when the QLQ-PBM had a larger mean difference relative to the SD, the EQ-5D questionnaire still reported larger mean change scores.

Discussion

This study developed three CS-PBMs from existing questionnaires HAQ, MSIS-29, and QLQ-C30 to provide evidence concerning comparability of CS-PBM–derived utility values with generic PBM–derived utility values. CS-PBMs had different mean utility values within a disease and did not report equal differences in mean utility values between diseases. The CS-PBMs in this study did not seem to exaggerate health problems, but rather reported higher mean values. Capturing comorbidities and along that line side effects of interventions appeared problematic for the HAQ-PBM, but not for the MSIS-PBM and the QLQ-PBM. The MSIS-PBM and the QLQ-PBM were more sensitive than the EQ-5D questionnaire to very mild impairment. The physical scale of the MSIS-29 questionnaire is known to be more sensitive in discriminating between clinical categories in MS than is the EQ-5D questionnaire. The MSIS-PBM, derived from the MSIS-29, also has better discriminatory properties.

Because the mean utility values of all three CS-PBMs were higher than those of generic instruments, it seems that a potential downward bias of a focusing effect may be smaller in size than the upward bias that results from a narrower scope of the condition-specific measures. This is most clearly seen in the performance of the HAQ-PBM, which is developed from the HAQ-Disability Index, which measures functional ability [24]. Consequently, the HAQ-PBM indicates the utility decrements associated with these functional (dis)abilities. In the HAQ-PBM, there is no dimension such as "pain" or "psychological state." Because pain is a frequently occurring symptom in arthritis, it is not surprising that the mean utility

Table 7 – The MSIS-PBM and the QLQ-PBM have increased sensitivity at the ceiling of the EQ-5D questionnaire.

Total sample size		EQ5D questionnaire = 1	EQ5D questionnaire <1	Worst state for which the EQ-5D questionnaire = 1
738	HAQ-PBM < 1	n = 7	–	21211
	HAQ-PBM = 1	–	n = 252	
1295	MSIS-PBM < 1	n = 99	–	33111222
	MSIS-PBM = 1	–	n = 2	
1505	QLQ-PBM < 1	n = 185	–	24334324
	QLQ-PBM = 1	–	n = 4	

EQ-5D, EuroQol five-dimensional; HAQ-PBM, Health Assessment Questionnaire-preference-based measure; MSIS-PBM, Multiple Sclerosis Impact Scale-preference-based measure; QLQ-PMB, Quality of Life Questionnaire-preference-based measure.

Table 8 – Comorbidities and side effects by PBM and the EQ-5D questionnaire.

	HAQ-PBM				EQ-5D questionnaire			
	Comorbidity		No comorbidity		Comorbidity		No comorbidity	
	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median
Diabetes	0.92 (0.08)	0.9	0.89 (0.10)	0.89	0.66 (0.24)	0.79	0.71 (0.16)	0.78
Hypercholesterolemia	0.90 (0.09)	0.88	0.89 (0.10)	0.89	0.57 (0.27)	0.65	0.72 (0.15)	0.78
Thyroid disease	0.90 (0.10)	0.89	0.89 (0.10)	0.89	0.60 (0.29)	0.71	0.72 (0.16)	0.78
Other cardiac disease	0.86 (0.05)	0.88	0.89 (0.10)	0.89	0.60 (0.21)	0.68*	0.72 (0.15)	0.78*
Psychiatric disorder	0.84 (0.13)	0.89	0.89 (0.10)	0.89	0.54 (0.30)	0.67*	0.72 (0.16)	0.78*
	MSIS-PBM				EQ-5D questionnaire			
	Comorbidity		No comorbidity		Comorbidity		No comorbidity	
	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median
Depression	0.61 (0.12)	0.59*	0.68 (0.14)	0.68*	0.54 (0.26)	0.64*	0.63 (0.26)	0.67*
Asthma	0.67 (0.14)	0.71	0.67 (0.14)	0.68	0.63 (0.22)	0.68	0.62 (0.26)	0.67
HBP	0.68 (0.14)	0.73	0.65 (0.13)	0.68	0.64 (0.25)	0.64	0.60 (0.26)	0.67
	QLQ-PBM				EQ-5D questionnaire			
	Side effects		No side effects		Side effects		No side effects	
	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median
Neurotoxicity†	0.76 (0.09)	0.76*	0.81 (0.09)	0.83*	0.56 (0.27)	0.65*	0.73 (0.25)	0.81*
Hair loss†	0.8 (0.08)	0.79	0.81 (0.1)	0.83	0.7 (0.23)	0.78	0.72 (0.26)	0.79
Nausea†	0.73 (0.13)	0.74*	0.81 (0.09)	0.83*	0.56 (0.31)	0.65*	0.72 (0.25)	0.81*
Other side effects†	0.79 (0.09)	0.8*	0.81 (0.09)	0.83*	0.66 (0.23)	0.69*	0.72 (0.26)	0.81*
	Infection		No infection		Infection		No infection	
Ear/nose/throat†	0.72 (0.11)	0.72*	0.81 (0.09)	0.83*	0.42 (0.32)	0.31*	0.73 (0.25)	0.81*

EQ-5D, EuroQol five-dimensional; HAQ-PBM, Health Assessment Questionnaire-preference-based measure; HBP, high blood pressure; MSIS-PBM, Multiple Sclerosis Impact Scale-preference-based measure; QLQ-PMB, Quality of Life Questionnaire-preference-based measure; WHO, World Health Organization.

* Significant difference between comorbidities/no comorbidities at ($P < 0.05$) Wilcoxon rank-sum test.

† WHO grade ≥ 2.

value of the early arthritis cohort as measured by the HAQ-PBM is much higher than the mean utility value of the generic instruments; any additional utility decrement besides functional disabilities, such as pain, is not captured directly, if at all. In the case of the HAQ, this result could have been anticipated on the basis of the fact that the HAQ-Disability Index aims to offer a unidimensional assessment of functionalities and does not attempt to measure other dimensions of health because these are captured by other instruments that are part of the minimum data set internationally agreed on. The unidimensionality of the HAQ caused some problems in the valuation task. Because all items aim to measure the same underlying latent variable (functional ability), they are highly related. OMEP-generated states have favorable statistical properties but do not consider the sensibility of the combination of item levels. Consequently, one health state in the valuation study consisted of the counterintuitive combination "able to get up from a chair" and "not able to get up from the toilet." This particular state caused confusion with some of the respondents.

The HAQ-Disability Index does not intend to form a comprehensive assessment of relevant disease-specific health outcomes in patients with rheumatoid arthritis, and therefore could be rejected as offering a suitable basis for the development of CS-PBMs. The large deviations in mean utility values presented in this study between the HAQ-PBM and the EQ-5D questionnaire support this view. More generally, it can be concluded that instruments with a narrow scope, often identifiable through inspecting items or dimensions, are unsuitable as a base for CS-PBMs used for resource allocation.

The perceived insensitivity of existing generic instruments is an important motive for developing CS-PBMs. In this study, sensitivity of the CS-PBM and the EQ-5D questionnaire was compared by investigating ceiling effects and discriminative ability of the instruments between patients with and without comorbidity or side effects. A ceiling effect found in the EQ-5D questionnaire for mild impairments was not found in the MSIS-PBM and the QLQ-PBM (Table 7). One reason for this difference may be the descriptive system of the questionnaires: the three-level system of the EQ-5D questionnaire might result in a lower likelihood of reporting problems than the four-level systems of the CS-PBMs. Nevertheless, using CS-PBMs did not result in an exaggeration of health problems on average when compared with generic instruments in this study. Rather, the mean utility value of the MSIS-PBM and the QLQ-PBM was higher than that of the EQ-5D questionnaire. This may be a reflection of the smaller range in obtainable utility values, which skews the average upward. Bad EQ-5D questionnaire health states reflect very poor health, which is perhaps not captured in the MSIS-PBM and the QLQ-PBM. Indeed, the negative range of utility values as produced for the EQ-5D questionnaire has rarely been reproduced for other instruments. The EQ-5D questionnaire, the MSIS-PBM, and the QLQ-PBM performed equally well in distinguishing patients with comorbidities/side effects from patients without it. Only the HAQ-PBM performed poorly in this aspect. Interestingly, the MSIS-PBM and the QLQ-PBM displayed equal discriminative ability as the EQ-5D questionnaire despite having a much

Table 9 – Discriminant validity.

	HAQ-PBM		EQ-5D questionnaire		N
	Mean	SD	Mean	SD	
DAS-28					
Remission	0.98	0.04	0.76	0.20	11
Low DA	0.90	0.08	0.70	0.25	15
Moderate DA	0.90	0.09	0.67	0.22	70
High DA	0.83	0.07	0.51	0.29	27
	MSIS-PBM		EQ-5D questionnaire		N
	Mean	SD	Mean	SD	
EDSS					
0	0.80	0.14	0.81	0.22	35
1	0.78	0.14	0.78	0.23	74
2	0.73	0.14	0.72	0.23	262
3	0.68	0.14	0.63	0.25	206
4	0.66	0.13	0.63	0.23	248
5	0.63	0.10	0.64	0.19	103
6	0.60	0.11	0.54	0.25	201
7	0.58	0.11	0.46	0.27	78
8	0.57	0.07	0.40	0.31	17
9	0.47	0.07	0.09	0.10	5
	QLQ-PBM		EQ-5D questionnaire		N
	Mean	SD	Mean	SD	
WHO					
0	0.83	0.11	0.75	0.25	356
1	0.83	0.10	0.76	0.24	304
2	0.80	0.11	0.69	0.24	96
3	0.71	0.10	0.37	0.27	27

DA, disease activity; DAS-28, Disease Activity Score-28; EDSS, Expanded Disability Status Scale; EQ-5D, EuroQol five-dimensional; HAQ-PBM, Health Assessment Questionnaire-preference-based measure; MSIS-PBM, Multiple Sclerosis Impact Scale-preference-based measure; QLQ-PMB, Quality of Life Questionnaire-preference-based measure; WHO, World Health Organization.

smaller total scale size due to a higher "floor" (i.e., the lowest attainable value).

Superiority of CS-PBMs compared with the EQ-5D questionnaire in regard to their discriminative ability was not demonstrated for the HAQ-PBM and equivalence was shown for the QLQ-PBM. The MSIS-PBM showed better discriminative properties than did the EQ-5D questionnaire in EDSS subcategories. With additional evidence on known-group differences, this could prove the MSIS-PBM to be a contribution to cost-utility analyses. The original preference-based questionnaire MSIS-29 was the only measure for which empirical evidence indicated better discriminative properties than the EQ-5D questionnaire in MS data sets.

While a CS-PBM may have desirable statistical properties, such as expressed in effect size or the ability to identify significant differences between groups with or without side effects, partly due to a small SD of mean values, these properties may not reflect the absolute size of differences in utility values between groups. This has consequences for quality-adjusted life-year computation. Imagine a new drug that reduces nausea from cancer treatments. Using the figures from Table 8, the population not having nausea would have a higher utility with an effect size (Cohen's *d* with pooled SDs) of 0.57 for the EQ-5D questionnaire but a larger 0.73 for the QLQ-PBM. The absolute difference, however, would be 0.16 for the EQ-5D questionnaire and 0.08 for the QLQ-PBM. An implication of these results is that if a CS-PBM is developed to increase sensitivity compared with the EQ-5D questionnaire, statistical sensitivity is not a sufficient criterion.

Rather than because of concerns about the sensitivity of an existing generic PBM, a CS-PBM may also be developed because a PBM was not administered in, for example, a clinical trial. In this case, one could also choose to use the variation in responses on a condition-specific measure to estimate what a generic utility instrument such as the EQ-5D questionnaire would have been had it not been absent, a process called mapping [25]. It is important to reflect on the question which strategy for deriving utilities from a disease-specific instrument is most appropriate. The main difference between mapping and constructing a CS-PBM is that the development of a PBM assigns population weights (via TTO) to the item levels of a questionnaire, while a mapping function assigns weights to the items that are dependent on the generic measure it aims to estimate. As such, issues with insensitivity of the generic instrument are not resolved when mapping a condition-specific measure onto a generic PBM. In our view, a well-conducted and validated mapping function may be preferred to the development of a CS-PBM, because it yields utility values that compare better to the more frequently used generic instruments used in other economic evaluations, but only under the following circumstances: 1) there is no empirical evidence for the insensitivity of the generic instrument, 2) only use of mean utility values is intended rather than subgroup analysis [26], and 3) the health status or disease subtype of the sample on which the function was estimated is comparable to the sample on which the function is applied [15].

Findings here underline that the TTO health state values as modeled from a fractional factorial design can differ from direct TTO valuations of those states. Often but not always an OMEP is applied to allow the estimation of TTO values for all theoretically possible health states from only a fraction of health states. This study adopted that technique but also valued directly a selection of states that were observed frequently in patients. Using these states in the estimation of the preference algorithm resulted in lower scores for at least some of these states (Fig. 1). These results suggest that discrepancies exist between modeled TTO values and directly observed TTO values for the most occurring health states, which may affect the validity of the measure. Little guidance is available for researchers who wish to design a valuation study for a CS-PBM by using state-of-the-art techniques, and so it is not surprising that practices vary and this deserves more attention to ensure that high-quality CS-PBMs are produced. Ideally, the process of constructing the CS-PBM is supported by the original developers of the questionnaires. This is relevant, for example, to avoid wild growth of value sets (e.g., for the QLQ-C30 now multiple value sets exist derived via mappings [15,27–29]), to further guarantee quality, and to offer support to users of the CS-PBM.

Table 10 – Responsiveness of utilities in non-Hodgkin sample.

Follow-up	Cohen's *d*		Mean change	
	QLQ-PBM	EQ-5D	QLQ-PBM	EQ-5D
Second treatment cycle	0.13	0.17	0.02	0.05
Fourth treatment cycle	0.02	0.08	0.00	0.02
Sixth treatment cycle	−0.09	−0.06	−0.01	−0.01
3-mo follow-up	0.33	0.22	0.03	0.06
6-mo follow-up	0.25	0.10	0.02	0.02
10-mo follow-up	−0.01	−0.09	0.00	−0.02
18-mo follow-up	0.00	0.19	0.00	0.04

QLQ-PMB, Quality of Life Questionnaire-preference-based measure.

Constructing and using a CS-PBM for the purpose of resource allocation could be considered when the following conditions are met: empirical evidence disproves the sensitivity of existing generic instruments, empirical evidence proves the superiority of the condition-specific measure from which the new PBM will be derived, and the derived CS-PBM is shown to be superior to the existing CS-PBM, not just in terms of statistical sensitivity but also in terms of absolute differences. The development of CS-PBMs is welcome from an academic point of view because it pushes methodological frontiers and introduces new data for comparing measures in a field where no gold standard PBM exists. Use in resource allocation of these instruments, however, is warranted only when the above-mentioned conditions are met. The introduction of PBMs that are specific to a certain disease has the presupposed merit of sensitivity to disease-specific effects of interventions, but this article shows that such an advantage is not necessarily achieved. Furthermore, the possible increase in sensitivity is traded off to the loss of comparability of absolute differences in utility values, which are most important for economic evaluations. It is argued here that without convincing empirical evidence on the insensitivity of a generic instrument, using a CS-PBM introduces confusion about the appropriate outcome measures in cost-utility analysis and health-care decision making.

Acknowledgments

We thank Prof. B. Uitdehaag from the multiple sclerosis centre of the VU Medical Centre in Amsterdam and Dr. J. Luime from the Netherlands Expert Centre for Work-Related Musculoskeletal Disorders, University Medical Center Rotterdam for sharing their expertise. Furthermore, we thank Mike Horton, from the University of Leeds, for his suggestions for the Rasch analysis and Mark Oppe for sharing his thoughts on the design of the TTO study. We owe gratitude to Bart Groenendijk for his aid with setting-up the computer-assisted experiment, to Ming Au for automating the data extraction, and to Fleur van de Wetering and Sandra de Vries for their assistance during the TTO exercise.

Source of financial support: This study was funded by ZonMW: the Netherlands organization for health research and development.

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at doi:10.1016/j.jval.2011.12.003 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] Dolan P. Modeling valuations for the EuroQol health states. Med Care 1997;35:1095–108.
[2] Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care 2002;40:113–28.
[3] Yang Y, Brazier JE, Tsuchiya A, Young TA. Estimating a preference-based index for a 5-dimensional health state classification for asthma derived from the Asthma Quality of Life Questionnaire. Med Decis Making 2011;31:281–91.
[4] Brazier J, Czoski-Murray C, Roberts J, et al. Estimation of a preference-based index from a condition-specific measure: the King's Health Questionnaire. Med Decis Making 2008;28:113–26.
[5] Brazier J, Tsuchiya A. Preference-based condition-specific measures of health: what happens to cross programme comparability? Health Econ 2010;19:125–9.
[6] Fryback DG, Lawrence WF. Dollars may not buy as many QALYs as we think: a problem with defining quality of life adjustments. Med Decis Making 1997;17:276–84.
[7] Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. Measuring and Valuing Health Benefits for Economic Evaluation. New York: Oxford University Press, 2007.
[8] Brazier JE, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ 2004;13:873–84.
[9] Kind P, Brooks R, Rabin R, eds. EQ-5D Concepts and Methods: A Developmental History. Dordrecht, The Netherlands: Springer, 2005.
[10] Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol 2003;30:167–78.
[11] Hobart J, Lamping D, Fitzpatrick R, et al. The Multiple Sclerosis Impact Scale (MSIS-29): a new patient-based outcome measure. Brain 2001; 124:962–73.
[12] Aaronson NNK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85:365–76.
[13] ten Klooster PM, Taal E, van de Laar MAFJ. Rasch analysis of the Dutch Health Assessment Questionnaire Disability Index and the health assessment questionnaire II in patients with rheumatoid arthritis. Arth Care Res 2008;59:1721–8.
[14] Ramp M, Khan F, Misajon RA, Pallant JF. Rasch analysis of the Multiple Sclerosis Impact Scale MSIS-29. Health Qual Life Outcomes 2009;7:58.
[15] Versteegh MM, Leunis A, Luime JJ, et al. Mapping QLQ-C30, HAQ, and MSIS-29 on EQ-5D. Med Decis Making 2011; (online first)
[16] Young TA, Yang Y, Brazier JE, Tsuchiya A. The use of Rasch analysis in reducing a large condition-specific instrument for preference valuation. Med Decis Making 2011;31:195–210.
[17] Mavranezouli I, Brazier JE, Young TA, Barkham M. Using Rasch analysis to form plausible health states amenable to valuation: the development of CORE-6D from a measure of common mental health problems (CORE-OM). Qual Life Res 2011;20:321–33.
[18] Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol 2007;46(Pt 1):1–18.
[19] Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health 2004;7:s22–6.
[20] Brazier JE, Roberts J, Platts M, Zoellner YF. Estimating a preference-based index for a menopause specific health quality of life questionnaire. Health Qual Life Outcomes 2005;3:13.
[21] Brazier J, Roberts J, Deverill M. The estimation of a preference based measure of health from the SF-36. J Health Econ 2002;21:271–92.
[22] Lamers LM, McDonnell J, Stalmeier PFM, et al. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ 2006;15:1121–32.
[23] Stolk EA, Busschbach JJ. Validity and feasibility of the use of condition-specific outcome measures in economic evaluation. Qual Life Res 2003;12:363–71.
[24] Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: dimensions and practical applications. Health Qual Life Outcomes 2003;1:20.
[25] Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ 2010;11:215–25.
[26] Versteegh M, Rowen D, Brazier J, Stolk E. Mapping onto EQ-5D for patients in poor health. Health Qual Life Outcomes 2010;8:141.
[27] Kontodimopoulos N, Aletras VH, Paliouras D, Niakas D. Mapping the cancer-specific EORTC QLQ-C30 to the preference-based EQ-5D, SF-6D, and 15D instruments. Value Health 2009;12:1151–7.
[28] McKenzie L, Van der Pol M. Mapping the EORTC QLQ C-30 onto the EQ-5D instrument: the potential to estimate QALYs without generic preference data. Value Health 2009;12:167–71.
[29] Crott R, Briggs A. Mapping the QLQ-C30 Quality of Life Cancer Questionnaire to EQ-5D patient preferences. Euro J Health Econ 2010;11:427–34.